专业社交媒体中的主题知识元抽取方法研究*

■ 林杰 苗润生 张振宇
同济大学经济与管理学院 上海 200092

摘要：［目的/意义］以汽车论坛例，提出一种针对专业社交媒体文本的主题知识元抽取方法。［方法/过程］首先，通过LDA模型提取出汽车论坛中文本的主题，并进行去重，形成主题列表；其次，基于融合主题特征的深度学习模型T-LSTM模型构建适于汽车论坛本文的情感分析模型；然后，通过计算各词汇在图模型TextRank中的重要性与各词汇的Word2Vec主题相似度，抽取情感关键词与关键句，用于对文本主题与情感倾向的解释与补充；最后，对上述方法进行集成，输出结构化的主题知识元。［结果/结论］实验结果中，抽取得到的主题知识元合格率达到69.1%，表明本文提出的主题知识元抽取方法，能够围绕知识主题较为准确地抽取知识元，实现知识的结构化转换。

关键词：主题知识元 主题抽取 长短期记忆神经网络 情感分析
分类号：G202
DOI：10.13266/j.issn.0252-3116.2019.14.012

1 引言

知识元，又称为知识单元、知识元组，是用于操作和管理知识的知识基元，是可以自由切分、表达、存取、组织、检索和利用知识的独立的知识单位[1]。主题知识元是知识元的一种表达形式，知识元中的元素包含知识主题词以及主题相关的关键信息。由于主题词能够准确反映知识元之间的各种隐含的有效关联，如等级种属关系、并列同一关系、簇类关系等，主题知识元是较为合适的知识元表达方式[1]。本文定义专业社交媒体为互联网用户用于分享和交换针对某一专业事物的意见、见解、经验和观点的内容生产与交换平台。专业社交媒体是一类特殊的社交媒体，其一般形式为专业论坛或专业社区，例如“汽车之家”“小米社区官方论坛”“虎扑NBA论坛”等，相比其他社交媒体如微博、脸书等，专业社交媒体中的语料内容性质专业且多为长文本[2]。本文针对专业社交媒体语料数量巨大、长短不一、创作随意性强、口语化的特点[3]以及知识操作与管理的使用的需求，提出一种以文本标题与文本内容为数据源、结构为“文本主题、主题情感倾向、主题关键词、主题关键句”的主题知识元。

主题知识元有机地结合了知识管理和现代信息技术，汲取了知识管理中隐性知识分类、知识提炼、知识应用等思想，采用大数据处理、文本挖掘、机器学习等技术，在众多领域拥有较高的应用价值。从海量文本中抽取主题知识元，实现了对知识内容本身的检索、自由操作与管理，同时完成了知识的控制单位从文档到主题知识元的转变，提高了知识检索与操作的效率与灵活性。利用主题知识元中主题之间的关联度，能够实现知识的重组与创造，以及对知识的量化与评价[4]。此外，在专业社交媒中，海量的用户评论语料蕴藏着丰富的用户创新知识。实现从专业社交媒体语料到主题知识元的抽取，能够提炼海量评论语料中的高价值信息，降低知识获取的难度与成本。专业社交媒体主题知识元中的情感倾向、关键词与关键句，能够为当前主题热度的测度与监控提供数据基础。专业社交媒体的主题知识元抽取是多种知识管理与创新活动的基础，企业能够利用主题知识元进行客户需求挖掘，与用户合作开展互动创新；政府与学术机构能够利用主题知识元开展社交舆情的

* 本文系国家自然科学基金面上项目“社交媒体中用户创新价值度测量模型及互动创新管理方法研究”(项目编号:71672128)和同济大学基本科研业务费专项资金项目“基于大数据的社交网络传播机理与模型研究”(项目编号：1200219368)研究成果之一。
作者简介：林杰(ORCID：0000-0002-5421-603X)，教授，博士，博士生导师；苗润生(ORCID：0000-0002-4784-1654)，博士研究生，通讯作者，E-mail：rmiao@tongji.edu.cn；张振宇(ORCID：0000-0002-4888-4023)，博士研究生。
收稿日期:2018-08-12 修回日期:2019-02-24 本文起止页码:101-110 本文责任编辑:杜杏叶

传播仿真研究,梳理社交舆情的主题脉络,监控突发舆情主题事件,为制定舆情治理措施与舆情调控策略提供依据。

2 相关研究

本文的创新点包括:①由于主题是专业社交媒体文本内容的思想核心,本文针对专业社交媒体文本,设计了以主题为中心的知识元结构,包含本文主题、主题情感与主题关键词句三方面信息;②针对专业社交媒体中文本用户评论语料数量巨大、用户知识水平参差不齐、文本长度长短不一、内容杂糅且低质、用词专业且口语化等特点[2-3],本文提出了抽取主题知识元的方法与技术路线,并进行了实验验证;③由于专业社交媒体语料数据量庞大,需保证主题知识元抽取的质量与速度,本文将深度学习技术引入到知识元的抽取中,近年来深度学习在文本挖掘领域应用的迅速发展,情感分析、文本相似度计算等任务取得了巨大进步,为主题知识元抽取的质量与速度提供了保证。

在知识元抽取研究领域,温有奎等[5]对知识元的内容进行了定义与分类,描述并实现了针对文献资源的抽取方案,其知识元的结构包括"类型、名称、内容"三种元素。然而该知识元结构单一,缺少对有价值信息的提炼,本文设计了囊括主题、情感与关键词句的知识元结构,不仅对知识内容进行了抽取,同时对文本包含的隐性知识进行了提炼,抽取得到了文本的主题与情感倾向等隐性知识 。姜永常[6]基于知识网格体系结构,描述了从文本实体层到语义层再到知识单元层的转换框架,从理论与技术层面构建了知识演化框架,但并未具体实现该框架。本文系统地对主题知识元的抽取方法进行了描述,同时通过实验检验了抽取方法的有效性。刘淼等[7]提出了一种针对文献资源的基于主题句的知识元抽取方法,通过计算句子之间的相似度,实现了句子级别的知识元抽取。然而主题句的抽取只考虑了单篇文档内的主题,本文在抽取主题时考虑了全局性、文档级别的主题元素,减少了抽取主题的冗杂度。

在抽取文本类别方面,上述研究的主要抽取对象均为学术文献资源,学术文献资源一般拥有明确的主题分类与关键词,抽取知识元时无须考虑重复抽取主题与关键词,本文针对专业社交媒体文本,构建了主题与主题关键词句的抽取方法。杨亮[8]针对新浪微博中的文本,提出了使用句子内信息与全局信息融合的方法,实现了对产品评论语料中的产品属性抽取,进行了基于认知思维模式的情感分析,形成了结构为"产品、产品属性、情感倾向"的知识元。然而针对微博等短文本语料的知识元抽取方法无法适应专业社交媒体中的长文本与专业性词汇,本文采用了基于深度学习的词嵌入方法支撑主题情感与关键词句的抽取,更好地适应了专业社交媒体的文本特点。Y. Yin 等[9]发现产品评论和其它附加信息(如用户信息和产品信息)对使用神经网络进行情感分析联合分类建模很有帮助,因此本文在其基础上,将文本与文本主题作为特征,使用长短记忆神经网络(LSTM)联合建模,提升了情感分析正确率。

3 研究方法

3.1 研究思路与框架

本文对主题知识元的定义:在专业社交媒体语料库 D 中,拥有 M 篇文章,主题知识元 u 是从文章 m 的标题 h 与内容 c 中抽取到的结构为 < 文本主题 t, 主题情感倾向 p, 关键词 k_w, 关键句 k_s > 的知识元,即 u:(t,p,k_w,k_s)。

本文的研究思路见图 1。具体如下:①从专业社交媒体中爬取用户评论文本,构建用户语料库;②运用 LDA 模型进行主题抽取,并合并重复主题,得到主题模型与全局主题列表 T,该 LDA 模型为后续主题情感分析与主题关键词句的抽取提供主题基础;③利用 LDA 模型对帖子主题极性标注,同时进行情感标注,构建基于 T-LSTM 的情感分析模型,输出情感倾向 p;④基于 TextRank 算法与 Word2Vec 主题词相似度算法,计算关键词句的加权重要度,从而实现关键词 k_w 与关键句 k_s 的抽取;⑤集成上述模型,训练并封装主题知识元 u 的抽取方法,并进行实验分析与验证。

3.2 主题抽取模型

本文首先通过训练 LAD 模型,在专业社交媒体的语料库 D 中,挖掘出合适数量的主题,得到主题列表 T。LDA 主题模型能够抽取得到专业社交媒体语料库中全局性的主题列表 2,将每篇文档单独输入 LDA 模型,从而得到语料库中每个文档所对应的主题。该模型为面向主题的情感分析与关键词句抽取提供了主题基础。

3.2.1 构建 LDA 主题抽取模型　LDA 模型的主要思想是找到文档在主题上的分布情况,以及主题词在主题上的分布情况,即每个文档对应一个或多个主题,每个主题拥有多个主题词,核心步骤[10]为:统计各个文

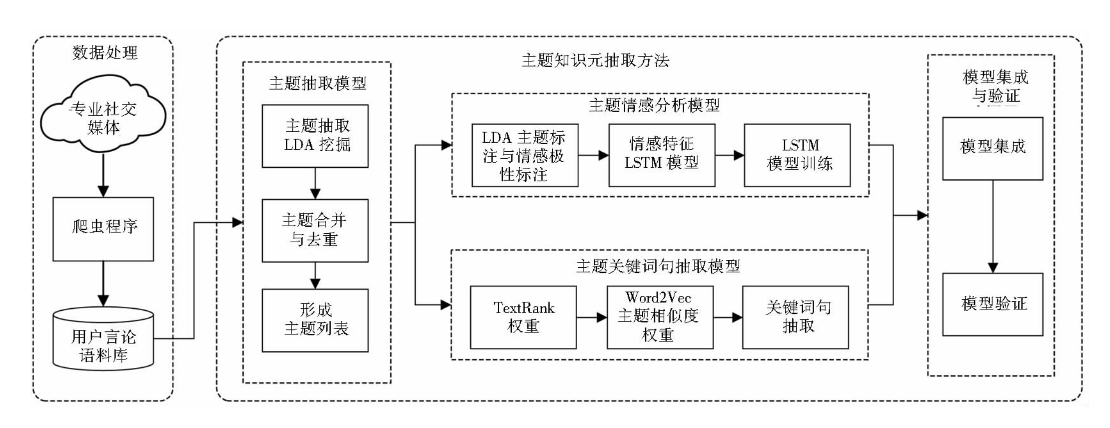


图1　主题知识元的抽取方法研究框架

档各个词的主题，得到文档主题分布 θ（见公式(1)），统计语料库中各个主题词的分布，得到 LDA 主题中主题词的分布 φ（见公式(2)）。

$$\theta_{mk}=\frac{n_{m,-i}^{k}+\alpha_k}{\sum_{s=1}^{K}(n_{m,-i}^{s}+\alpha_s)}\qquad \text{公式 (1)}$$

$$\varphi_{kt}=\frac{n_{k,-i}^{t}+\eta_t}{\sum_{s=1}^{K}(n_{k,-i}^{s}+\eta_s)}\qquad \text{公式 (2)}$$

公式(1)、(2)中，K 为主题的个数；α 为 θ_m 分布的超参数，表示主体之间的相对强弱，是一个 K 维向量，α_k 是 α 的第 k 个元素；η 为 φ_k 分布的超参数，是一个 T 维向量，T 为词典大小。公式(1)中的 $n_{m,-i}^{k}$ 是第 m 篇文章中分配到主题 k 的单词个数，不包含当前单词 i。公式(2)中 $n_{k,-i}^{t}$ 是第 k 个主题中分配到单词 t 的数，不包含当前单词 i。

主题分布 θ 中选取分布最高前 n 个主题作为初始主题列表 T_0，主题词的分布 φ 中每个主题所对应的分布最高前的 m 个主题词作为主题列表中的主题词 t。

3.2.2　主题去重　LDA 抽取得到的主题中会出现主题重复、冗余的情况，本文在得到初始主题列表 T_0 后，通过计算主题相似度的方法，进行主题去重，得到语料库 D 的主题列表 T。

计算两个主题是否重复或冗余，需要首先计算两个主题的相似度。抽取两个主题的前 W 个词，分别为集合 A 与集合 B，然后计算两集合的 Jaccard Similarity，公式如下：

$$J(A,B)=\frac{|A\cap B|}{|A\cap B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}\qquad \text{公式(3)}$$

遍历主题列表 T 中的主题，按照公式(3)计算两个主题间的相似度；将每个主题对的相似度 $J(A,B)$ 与给定的阈值(Jaccard Thresh)相比较，并记录下所有大于阈值的主题对，最后通过并查集(Disjoint-Set)方法进行合并，得到主题列表 T。

在主题去重后，主题分布更加合理，人工命名主题的工作量会相应减少，从而提高抽取速度与质量，最终抽取出属于语料库 D 的主题列表 T 以及相应的主题词 t。

3.3　主题情感分析模型

了解用户对文本中主题上的情感倾向，是知识元抽取的关键任务之一，针对专业社交媒体文本，该任务能够帮助知识使用者了解用户需求、量化用户评价等[11]。本文在 LDA 模型抽取主题的基础上，采用 LSTM 模型计算用户发表文本(帖子)的情感倾向。由于模型自身的递归特性[12]，本文在 LSTM 模型的训练过程中，使用主题标签与情感标签共同监督。共同监督一方面能够围绕文本主题抽来取情感倾向；另一方面，能够利用主题与情感倾向之间的相关性，提升情感分析的准确率。

3.3.1　主题与情感的相关性　在专业社交媒体中，用户发表的言论是以帖子的形式来表现的，由于帖子文本长短不一、类别鱼龙混杂，其情感属性难以把握。然而专业社交媒体中用户的帖子一般带有鲜明的主题，这些主题与情感倾向往往具有相关性[13]，以汽车论坛为例，帖子包含“买车晒车”“故障与维修”“配置对比”等主题。将由 LDA 模型得到各帖子的主题，与各帖子的情感倾向进行相关性分析后，认为帖子主题与帖子所包含的情感倾向，有较强的相关性。各主题下的情感倾向统计结果见表 1。

由表 1 可知，买车晒车主题的帖子多为正面情感；涉及到故障、异常与维修描述的帖子多为负面情感；而活动与社交帖子、其他类别的帖子例如二手交易帖，大多数不具有明确的情感倾向。从统计数据可知，帖子主题是一个强相关变量，将其输入 LSTM 模型作为特征进行学习，能够一方面使情感倾向结果贴近帖子主题，另一方面到提升模型分类效果的作用。

表 1 各主题下帖子的情感倾向统计

（单位：条数）

帖子目的性主题	负面	正面	中性	合计
1. 购车价格与程序	182	132	1 248	1 562
2. 汽车配置对比评价	341	451	1 315	2 107
3. 汽车改装讨论	176	342	1 358	1 876
4. 汽车保养讨论	269	153	126	548
5. 故障、异常与维修	1 321	84	110	1515
6. 使用求助	105	262	2 026	2 393
7. 买车晒车	58	703	253	1 014
8. 轮毂轮胎讨论	40	67	336	443
9. 活动与社交	2	9	308	319
10. 其他	44	25	154	223
合计	2 538	2 228	7 234	12 000

3.3.2 融合主题特征的 LSTM 情感分类模型 本文将在 LSTM 模型中引入事先通过 LDA 方法获取的主题信息的模型，命名为主题增强的 LSTM 情感分类模型（T-LSTM）。T-LSTM 的主要思想是：利用 LSTM 中各隐藏层的递归性，将 LDA 模型的主题词信息作为输入序列的后续的时间节点（Time-step）输入模型，然后利用样本在该主题上的情感倾向标注进行训练，通过学习情感倾向与主题信息的相关性，提高输出该主题上的情感倾向 p 的准确率。

T-LSTM 模型的整体结构见图 2，共包含 3 层网络，自下到上分别是 Embedding 层、LSTM 层（见图 3）、MLP 层。

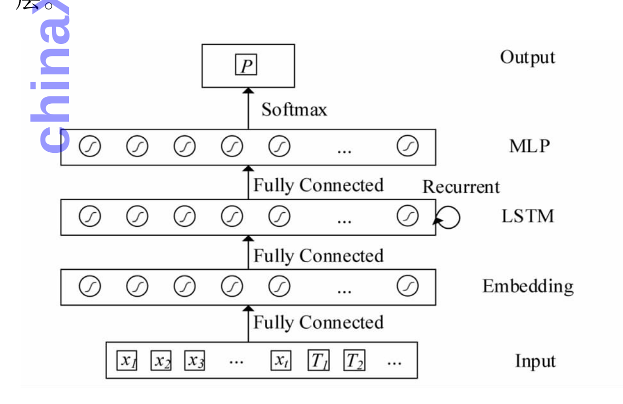


图 2 T-LSTM 网络结构

第一层，词嵌入层（Embedding）。词嵌入层位于整个模型的最底部，作用是对经过 One-hot Vector 处理的词向量进行降维，从而减少模型的复杂度。词嵌入的输出 y_i 作为学习模型 $g:y\rightarrow z$ 的输入，已知任务 g 中所对应 z_i 值。通过样本数据 $\{(x_i,z_i)_{i=1}^{N}\}$ 训练得到学习模型 $k:x\rightarrow z$，即 $z=g(f(x))$，该过程中的模型 $y=f(x)$ 即为词嵌入的模型。

第二层，T-LSTM 核心网络层。此处使用 T 代表主题向量，P 代表情感倾向向量，整个 LSTM 层的输入是不同时间节点 t 的词向量（$x_1, x_2, \cdots, x_t$），以及后续的主题词向量（$T_1, T_2, \cdots$），输出为情感倾向 P 所对应的向量。引入主题信息作为特征后，LSTM 核心层的结构见图 2。图 3 中每个节点代表一层包含了一个记忆块的隐藏层，每个记忆块的输入是上一层的输出与改成的输入，即帖子文本层的输出会作为主题特征层的输入，最后的输出层既包括文本信息又包括主题信息。图 3 中 W_1、W_2 是输入与输出向量的权重矩阵，W_P 分别是情感倾向 P 在隐藏层之间的权重矩阵。

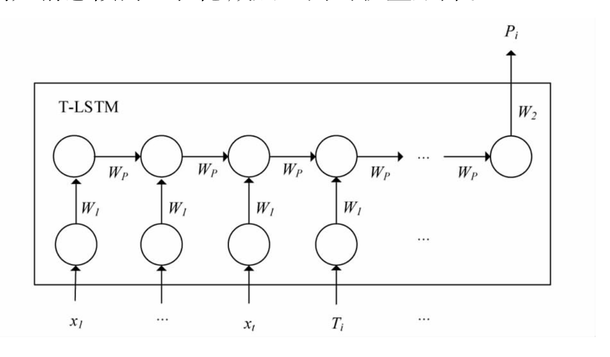


图 3 LSTM 层的网络结构

第三层，多层感知机网络（Multi-layer Perception，MLP）。将第二层得到的主题向量与情感倾向向量输入到 MLP 层，MLP 层输出的向量通过 Softmax 层得到情感倾向标签的概率（P_k）[12]，k 代表情感倾向，在模型训练得到的参数 μ 的条件下，目标概率可分别表述为：

$$p(P_k \mid x,\mu) = \frac{e^{W_k^P x + b_k^P}}{\sum_{i=1}^{|P|} e^{W_i^P x + b_i^P}} \qquad \text{公式(4)}$$

公式（4）中 e 为自然对数底，x 向量表示上一隐藏层节点输出的值，（$W_k^P x + b_k^P$）代表通过感知机层权重 W_k^P 与截距 b_k^P 计算得到的未归一化的概率。

假定训练样本为 M，模型中的节点数为 S，那么在训练时定义损失函数为：

$$L(\mu) = \frac{1}{M}\sum_{s\in M}\sum_{x\in s} l\{P_k = j\}\times \log p(P_k \mid x,\mu) + \alpha|(|\mu|)|^2$$

公式（5）

公式（5）中 $l\{P_k=j\}$ 表示如果 $P_k=j$ 成立，则 l 的值为 1，否则为 0；$\alpha|(|\mu|)|^2$ 为损失函数中的惩罚项，μ 为模型中训练得到的参数，α 为惩罚系数，取值在［0，1］之间。

本模型的训练采用 Adam 算法[12]，利用梯度的一阶矩估计和二阶矩估计动态调整每个参数的学习率。此处使用了 Dropout 技术来防止模型过拟合，适的并采用 mini-batch 的方法进行训练。训练完成后，对模型

进行序列化保存,最后利用该模型输出主题知识元中的情感倾向 p。

3.4 主题关键词句抽取模型

LDA 所抽取到的主题是所有文档层面得到的全局性主题,对于单个文档来说,其自身的主题往往无法与 LDA 得到的主题词一一对应,会出现主题词冗余的问题。而文档级别的关键词句不同于 LDA 主题,是对文档本身关键信息的抽取,抽取得到的关键信息比 LDA 主题词粒度更细,且关键词来源于该文档本身。本文综合文档级别关键词与 LDA 主题词抽取算法,在不偏离大主题的前提下,抽取文档关键词与关键句,提供对 LDA 主题以及情感倾向的解释。因此,本文一方面使用 TextRank 算法计算单一文档中词、句的重要度,考虑文档级别的关键词句抽取;另一方面使用 Word2Vec 算法计算文档中词句与文档主题词的相似度,考虑抽取得到的关键词句在一定程度上与 LDA 主题词相一致。最终,通过加权的方法计算综合重要度,从而选取重要程度最高的词与句,作为该帖子的关键词 k_w、关键句 k_s。

3.4.1 基于 TextRank 的关键词句重要度计算 本文采用 TextRank 来抽取帖子中的关键信息,其基本思想是:首先将文本分割成若干组成单元(单词、句子),之后建立图模型,采用投票机制对文本中的成分进行排序[14]。该算法的优势在于不需要事先对多篇文档进行学习训练,仅基于单个文档的信息即可完成关键词的提取与文摘,完成过程简洁有效。

TextRank 模型的结构为一个有向有权图 $G=(V,E)$,由点集合 V 和边集合 E 组成,其中 E 是 $V \times V$ 的子集。有向图中任两点 V_i, V_j 之间边的权重为 w_{ji}。对于任意一个给定的点 V_i, $IN(V_i)$ 表示指向该点的点集合, $OUT(V_i)$ 表示点 V_i 指向的点集合。则点 V_i 的得分 $WS(V_i)$ 定义如下:

$$WS(V_i)=(1-d)+d \times \sum_{V_j \in IN(V_i)} \frac{w_{ji}}{\sum_{V_k \in OUT(V_j)} w_{jk}} WS(V_j) \quad \text{公式 (6)}$$

公式(6)中 d 为阻尼系数, 其取值范围为 0 到 1, 表示从图中某一特定点指向其他任意点的概率,默认取值为 0.85。上式中 $WS(V_j)$ 是点 V_j 的得分,该式通过递归迭代的方式进行计算,因此每个点的得分需要赋予一个随机初始值。

关键词抽取的目标是从给定的文本中自动抽取出若干有意义的词语或词组。其步骤包括:①将文本 M 按照句子 S_i 进行分割,用 $k_{i,j}$ 表示句中的词汇;②构建图 $G=(V,E)$,其中 V 为包含了 $k_{i,j}$ 的集合,E 为利用共现窗口构建两点间的边;③根据公式(6),迭代计算各节点权重,直至收敛;④倒序排序各点权重;⑤提取最重要的 N 个单词,如果形成相邻词组,则组合成多词关键词。

同理,在上述步骤中,将词替换为句子,抽取得到关键句。

3.4.2 基于 Word2Vec 的关键词句主题相似度计算 Word2Vec 模型是由谷歌提出的词向量模型,它尝试通过分析一个词的邻词(也称作语境)来确定该词的含义。因此通过训练 Word2Vec 模型,能够使用词向量之间的距离来表示词语的语义相似性。

本文通过训练 Word2Vec 模型,得到所有语料中的词汇的词向量,然后利用词向量,计算文本中词、句与主题词的相似度。词语相似度采用余弦相似度来计算,a,b 表示两个词汇的词向量:

$$cos\theta = \frac{ab}{||a||\,||b||} \quad \text{公式 (7)}$$

计算文档中每一个词与该文档的 LDA 主题中的主题词的相似度,取最高的主题词的相似度作为该词的主题相似词相似度:

$$cos\theta_i = Max_{j \in Topic_{words}}(cos\theta_j) \quad \text{公式 (8)}$$

然后以该贴单词集合为主体,进行归一化处理,得到单词的主题相似度:

$$sim_i = \frac{count_i * cos\theta_i}{\sum_{k \in All_{Words}} cos\theta_k} \quad \text{公式 (9)}$$

其中,$count_i$ 代表单词 i 出现的次数,k 是该文档中出现的单词。

对于句子与主题相似度的计算,本文采用将句子中与主题词中相似度最高的 m 个(默认值 $m=3$)词相似度之和,并根据文档所有句子相似度之和进行归一化,得到的值作为句子的主题相似度。

3.4.3 加权计算关键词句重要度 综合使用主题相似度与 TextRank 重要度,来确定文档中词句的重要度,文档中每个词的重要度使用公式 10 计算。

$$I_i = w \times Sim_i + (1-w)\,TextRank_i \quad \text{公式(10)}$$

其中 w 代表主题相似度所占的权重,取值在[0,1]之间,$TextRank_i$ 表示该词的 TextRank 重要度,Sim_i 表示该次的主题相似度。同理,文档中句子的重要程度也可通过公式(10)计算。

最后,将文档词汇与文档句子按照加权的重要度 I 倒序排列,截取权重最高的 T 个单词作为关键词 k_w,截取权重最高的 N 个句子得到关键句 k_s。

4 实验结果与分析

汽车产品是最复杂的工业产品之一,汽车行业有着庞大的技术体系、多变的市场需求、高昂的研发与制造成本[15]。此外,汽车产品价值较高、与人们生活息息相关,是各类工业产品中,普遍关注的重要产品。因此本文以汽车论坛为例,开展专业社交媒体中的主题知识元抽取实验。

4.1 汽车文本爬取

本文通过编写基于 Scrapy 的爬虫程序,抓取汽车之家论坛中的汽车评论帖子,选取了 10 个热门车型论坛进行爬取,包括迈腾论坛、雅阁论坛、凯美瑞论坛等,爬取内容包括帖子标题、正文内容、配图文本等信息,时间范围为从 2016 年 9 月至 2017 年 9 月。删除内容为空或 5 个字符以下的帖子,删除内容过长的灌水帖,即字数超过 500 字、却只包含不超过 20 个不同字符的帖子。共爬取 10 万余条汽车评论帖子。

4.2 文本主题抽取

4.2.1 训练 LDA 模型并输出主题列表 本文采用 Python 中"Topic Modeling with Latent Dirichlet Allocation"库,实现 3.2.1 节中所描述的算法过程。首先对帖子进行预处理,去除常用词、地名、品牌等名词。然后对模型的参数进行选择,主题个数 $K=20$,狄利克雷分布超参数 $\alpha=0.1$, $\eta=0.01$,迭代次数 $iterations=100$,并使用预处理后的帖子训练 LDA 模型后,得到初始主题列表 T_0。在通过 LDA 获得主题列表 T_0 后,对其中每个主题下的主题词进行同义词合并、无意义词剔除的处理,例如,发动机、引擎合并为发动机,轮胎、车胎合并为轮胎等。然后根据 3.2.2 节中的主题去重,此处设置每个主题取前 $W=20$ 个词进行相似度计算,主题合并的相似度阈值 $t=0.1$,即相似度超过 0.1 的主题将进行合并,得到主题列表 T。在运行去重模型后,原先 LDA 主题列表中的 20 个主题合并为 10 个主题,如表 2 所示,其中"主题"一列是根据 LDA 算法得到的分布最高的 10 个主题词进行人工命名得到的主题名称。

表 2 去重后的主题列表

主题序号	主题	分布最高的 10 个主题词
1	购车价格与程序	优惠、销售、贷款、价格、提车、落地、购置税、保险、订车、加价
2	汽车对比评价	配置、动力、运动、油耗、空间、变速箱、后排、内饰、落地、安全
3	汽车改装讨论	改装、导航、影像、大灯、升级、轮毂、、安装、雷达、原车、疝气
4	汽车保养讨论	机油、保养、美孚、机滤、清洗、滤芯、节气门、火花塞、防冻液、空调
5	故障、异常与维修	异响、声音、追尾、刹车、抖动、问题、变速箱、顿挫、熄火、发动机
6	使用求助	大神、求助、请教、帮忙、进来、告知、指教、车友、请问、指导
7	买车晒车	实体店、作业、提车、版主、认证、颜色、好看、推荐、内饰、系统
8	轮胎讨论	轮胎、轮毂、胎压、影响、备胎、原厂、补胎、米其林、定位、磨损
9	活动与社交	猜车、活动、微信、车友会、咨询、交流、加入、软件、支持、音乐
10	其他	删除、本楼、管理员、精华、领先、论坛、自动、帖子、喜欢

4.2.2 运用 LDA 模型输出各文档主题 使用已经训练完毕的 LDA 模型,反向运行,输出每个帖子的主题分布,部分文档获取的主题如表 3 所示,表中展示了概率最高的 2 个主题编号与主题名称。

表 3 各帖子抽取得到的主题展示

帖子序号	帖子名称	主题 1	主题 2
1	一汽大众迈腾 B7L 发动机设计缺陷 导致顶气门	5. 故障、异常与维修	2. 汽车对比评价
2	第一次和迈腾 B8 的亲密接触	2. 汽车对比评价	7. 买车晒车
3	B7 迈腾近四年,些许问题请教老司机 ~	6. 使用求助	4. 汽车保养讨论
4	【吃胎更新】一代 18000 公里车况解说	8. 轮胎讨论	4. 汽车保养讨论
5	别人都提新款了,我提老款 1.8 舒适	7. 买车晒车	1. 购车价格与程序
共 12 000 帖	…	…	…

4.3 主题情感抽取

本文同时使用 T-LSTM 模型、LSTM 模型与 SVM 模型进行情感分析实验,并对实验结果进行对比分析。

4.3.1 人工标注数据集 采用 T-LSTM 模型进行情感分析,需要高质量标注的主题情感倾向标签,这些标签应该围绕文本本身的主题进行标注。

本文在语料库中选取 12 000 篇帖子,然后使用 LDA 模型提取每篇帖子的主题,主题属于上述主题列表中的 10 类主题。然后组建 8 人标注小组,依据该文本主题相关的情感倾向进行分工标注。此外,为了保障标注质量,标注工作将进行交叉校验,即每篇帖子会有 2 人进行交叉标注,对标注结果不同的帖子重新进

行标注。具体标签的数量分布见表1。将帖子按照2:1的比例划分为训练集与测试集,分别为8 000篇与4 000篇。

4.3.2 模型超参数选取 采用训练集的5倍交叉验证(5 – fold Cross-Validation)来选取模型的超参数,所选出的超参数也将用于下面的实验。其中T-LSTM模型与LSTM模型选取相同的超参数,其中词典的数量w的取值范围为(5 000,20 000),搜索间隔为1 000;词向量的维度d取值范围为(50,200),搜索间隔为10;LSTM隐藏节点数H_l(100,1 000),搜索间隔为100,漏码率dropout,记为$drpt$取值为(0.5,0.9)搜索间隔为0.1;为减少模型计算复杂度,MLP的隐藏层数为1,隐藏层节点H_m的取值范围为(50,200)。采用网格搜索(Grid Search)方法选择使得平均准确率最优的一组,该组数据见表4。此外,SVM模型的正则化常数C取值为1.0。

表4 T-LSTM超参数取值说明

参数	参数说明	参数值
w	词的数量	12 000
d	词向量维度	128
H_l	LSTM隐藏节点数	200
$drpt$	漏码率 dropout	0.8
H_m	MLP隐藏节点数	160

4.3.3 实验结果分析 该实验是三分类问题,在包含4 000篇帖子的测试集中,"正面""负面"以及"中性"标签的数量分别为845、743、2 412。在训练过程中使用了不同样本数的训练集,其效果见图4,可见在训练集大小为4 000时,T-LSTM的效果开始优于LSTM与SVM。由于T-LSTM模型的复杂度高于其他两个模型,因此在训练集足够大时有相对优势。根据图4可知,在训练集大小为8 000时,T-LSTM、LSTM与SVM在测试集上的准确率分别为84.9%、82.6%与80.4%。T-LSTM相比LSTM与SVM,正确率分别提高了额2.3%与4.2%。

表5、表6、表7分别是LSTM、SVM、T-LSTM模型在测试集上的混淆矩阵,其中标签"0""1""2"分别代表"正面""负面""中性",其实际数量分别为845、743、2 412。通过混淆矩阵能够清晰地了解各模型预测正确与错误的情况。相比之下,T-LSTM模型混淆矩阵中,标签"0""1""2"下,预测正确的数量分别为621、591、2 161,均高于其他两个模型相应标签预测正确的数量。

根据上述分析可知,T-LSTM模型在融入主题特征以及改进LSTM结构的情况下,在样本集充足的情况下,能够发挥LSTM模型处理序列数据的优点,同时通过将主题信息输入模型,提高了帖子在主题方向上情感分析的准确率。

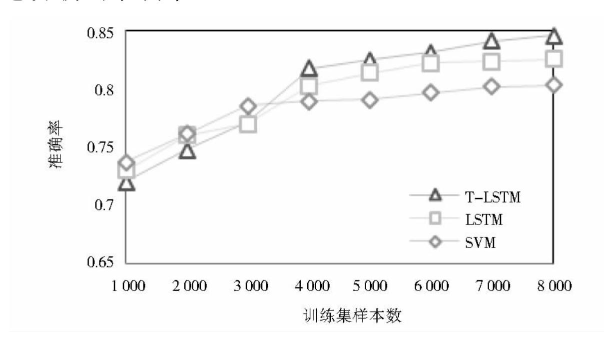


图4 不同样本数训练集下的模型准确率

表5 LSTM模型混淆矩阵

频数		预测值 0	预测值 1	预测值 2	合计
实际值	0	589	74	182	845
	1	36	572	135	743
	2	154	115	2 143	2 412
	合计	779	761	2 460	4 000

表6 SVM模型混淆矩阵

频数		预测值 0	预测值 1	预测值 2	合计
实际值	0	585	74	186	845
	1	57	557	129	743
	2	153	185	2 074	2 412
	合计	795	816	2 389	4 000

表7 T-LSTM混淆矩阵

频数		预测值 0	预测值 1	预测值 2	合计
实际值	0	621	32	192	845
	1	35	591	117	743
	2	132	119	2 161	2 412
	合计	788	742	2 470	4 000

4.4 主题关键词句抽取

主题关键词句的抽取将以两篇帖子为例(帖子内容可见表8中帖子1与帖子2),展示关键词句抽取的过程与结果。

4.4.1 TextRank重要度计算与Word2Vec主题相似度计算 TextRank重要度计算:通过编写Python程序,按照3.4.1节中的算法,对样本汽车帖子文本进行实验,模型中参数共现窗口长度$K=6$,关键词个数$T=20$,关键词最小出现次数为1。基于TextRank的关键词计算结果见表9,关键句计算结果见表10。

表 8 主题知识元抽取结果

帖子	帖子标题 h	帖子内容 c	输出	
帖子 1	一汽大众迈腾 B7L 发动机设计缺陷 导致顶气门	大众迈腾 2013 年 8 月 2 号购买!行驶 48 000 公里,昨天下班正常开车,突然车剧烈抖动,失去动力熄火,之后无法启动,还好当时路上车不多,如果是在快速路 高速公路,那后果不堪设想,车辆拖走,经检查由于张紧器问题,正时链条跳齿,导致顶气门!致电厂家 400 以已经过保为由拒绝一切赔偿,将近 30 万的车,跑了 3 年 4.8 万公里出此问题,车辆质量上有严重问题,不能因为厂家产品设计问题,过保就要用户自己买单。我不是黑大众,大众实在是太让人伤心了,完全不顾用户安全,烧机油我忍了,变速箱我忍了,异响我忍了,在长春冬季后车门无法打开我忍了,这回又来发动机,大众啊,迈腾啊,你真此乃神车啊!我服了,真的服了,开了几年我基本都快会修车了。以上句句属实。	文本主题 t	5. 故障、异常与维修
			情感倾向 p	2. 负面
			关键词 k_w	问题、发动机、抖动异响、无法
			关键句 k_s	行驶 48000 公里,昨天下班正常开车,突然车剧烈抖动,失去动力熄火,之后无法启动,还好当时路上车不多,如果是在快速路 高速公路,那后果不堪设想,车辆拖走,经检查由于张紧器问题,正时链条跳齿,导致顶气门。
帖子 2	第一次和迈腾 B8 的亲密接触	同事要买车,下班赖在我车上,要坐我载他去看车。也真是没有办法了,冲着他说的请客吃晚饭就陪他跑一跑吧。抓紧时间把这三家店跑了一圈。同事在考虑英朗、速腾、宝来、朗逸和凌渡这几个车,依我看肯定速腾和凌渡这两个好点,看起来档次都不一样。这一圈逛下来我最感兴趣的车就新迈腾,这车上市很久了,我这还是第一次零距离跟它接触。车子外观空间都不错,给外观点赞,尤其内饰给我的感觉很好,坐进车里第一感就是精致加豪华。展厅摆的是 330 豪华型的迈腾,落地不到三十万,内饰的整体表现绝对符合这个价位。多处用的是软性材质,摸上去质感不错,真皮座椅柔软手感好。……	文本主题 t	2. 汽车对比评价外
			情感倾向 p	1. 正面
			关键词 k_w	内饰、落地、同事、空间、后排
			关键句 k_s	展厅摆的是 330 豪华型的迈腾,落地不到三十万,内饰的整体表现绝对符合这个价位。
…	…	…	…	…

表 9 关键词的 TextRank 重要度展示

帖 1	词 1	词 2	词 3	词 4	词 5
词语	问题	无法	忍	用户	发动机
权重	0.029	0.024	0.022	0.021	0.019
帖 2	词 1	词 2	词 3	词 4	词 5
词语	内饰	同事	看	落地	质感
权重	0.031	0.019	0.018	0.017	0.013

表 10 关键句的 TextRank 重要度展示

帖 1	句 1	句 2
词语	行驶 48 000 公里,昨天下班正常开车,突然车剧烈抖动……	我服了,真的服了,开了几年我基本都快会修车了。
重要度	0.204	0.189
帖 2	句 1	句 2
词语	展厅摆的是 330 豪华型的迈腾,落地不到三十万……	没有试驾,也就随便这么一看,不过对迈腾的……
重要度	0.109	0.106

Word2Vec 主题相似度计算:本文采用 Python 中 Gensim 库训练 Word2Vec 模型,设置模型训练参数词向量的维度 *size* = 100,学习率 *alpha* = 0.05,词最低频率 *mincount* = 3,训练的窗口大小 *window* = 5,将 12 余万篇帖子分词后输入模型,训练后得到所有词汇的词向量。然后,获取每篇帖子的 LDA 主题,根据 3.4.2 节中的方法计算帖子中词语的主题相似度。例如,帖子 1 的主题为:“5. 故障、异常与维修”,其主题词包括:“异响、声音、抖动、追尾、刹车、问题、变速箱、顿挫、熄火、发动机”;帖子 2 的主题为:“2. 汽车对比评价”,其主题词包括:“配置、动力、运动、油耗、空间、变速箱、后排、内饰、后排、安全”。帖子 1 中“抖动”“异响”“问题”“变速箱”“发动机”等词在本帖子 LDA 主题词也出现,因此其主题相似度更高,同理可得帖子 2 的结果。基于 Word2Vec 的关键词主题相似度结果见表 11,关键句主题相似度结果见表 12。

表 11 关键词的 Word2Vec 主题相似度展示

帖 1	词 1	词 2	词 3	词 4	词 5
词语	抖动	异响	问题	变速箱	发动机
相似度	0.025	0.025	0.013	0.013	0.007
帖 2	词 1	词 2	词 3	词 4	词 5
词语	空间	内饰	后排	外观	落地
相似度	0.022	0.019	0.019	0.019	0.013

表 12 关键句的 Word2Vec 主题相似度展示

帖 1	词 1	词 2
词语	我不是黑大众,大众实在是太让人伤心了,完全不顾……	行驶 480 00 公里,昨天下班正常开车,突然车剧烈抖动……
相似度	0.221	0.149
帖 2	词 1	词 2
词语	车子外观空间都不错,给外观点赞,尤其内饰给我的……	展厅摆的是 330 豪华型的迈腾,落地不到三十万……
相似度	0.137	0.125

4.4.2 重要度加权计算 在 TextRank 重要度与 Word2Vec 主题相似度计算的基础上,根据公式 10 计算加权的关键词、句重要度,其中主题相似度权重 w 设为 0.5,关键词、关键句的计算结果分别见表 12 与表

13。

表 13 中,帖子 1 中的加权关键词为“问题”“发动机”“抖动”“异响”“无法”,相比单独使用两种方法,“问题”与“发动机”两词的权重弄增加,“抖动”与“异响”两个主题相似度高的词入选为关键词,与表 8 对比,其针对帖子 1 内容的入选关键词更加合理。帖子 2 中的在加权计算后,高主题相似度的“空间”“内饰”“落地”入选关键词,“看”与“质感”被排除,与表 8 对比,其关键词更加贴近帖子 2 的主题。

表 13 加权的关键词重要度展示

帖 1	词 1	词 2	词 3	词 4	词 5
词语	问题	发动机	抖动	异响	无法
重要度	0.021	0.013	0.013	0.013	0.012
帖 2	词 1	词 2	词 3	词 4	词 5
词语	内饰	落地	空间	同事	后排
重要度	0.025	0.015	0.011	0.009	0.009

关键句抽取使用基于 Word2Vec 的主题相似度算法与 TextRank 算法,加权计算每篇帖子中的关键句的重要程度,也拥有相同效果,计算结果见表 14。

表 14 加权的关键句重要度展示

帖 1	句 1	句 2
词语	行驶 48 000 公里,昨天下班正常开车,突然车剧烈抖动……	我不是黑大众,大众实在是太让人伤心了,完全不顾……
重要度	0.176	0.110
帖 2	句 1	句 2
词语	展厅摆的是 330 豪华型的迈腾,落地不到三十万……	车子外观空间都不错,给外观点赞,尤其内饰给我的……
重要度	0.115	0.069

通过上述流程,对语料库中所有帖子进行关键词句抽取,随机选择其中 2 000 篇帖子进行人工校验,使用单一 TextRank 进行抽取的关键词句合格条数为 1 402条,合格率为 70.1%;融合主题相似度的加权算法抽取到的关键词与关键句通过检验的数量为 1 562 条,合格率为 78.1%,合格率提高 8%。

4.5 主题知识元抽取模型集成

将文本主题 t、情感倾向 p、关键词 k_w、关键句 k_s 的抽取方法进行集成,使用集成模型将语料转换为结构化的主题知识元进行储存,以便于在产品创新时对所需知识进行检索与使用。

首先,初始化集成模型,将文本语料库 D 输入 LDA 与 LSTM 模型进行训练,得到训练完成的模型文件。然后,编写 API 接口程序来加载 LDA 与 LSTM 的模型文件、输出模型结果、实现关键词句抽取过程。最终,输入单个文本标题与本文内容,输出结构化的主题知识元。该 API 实现了从本文标题 h、本文内容 c、主题列表 T 到主题知识元的映射,即 $f_u:(h,c,T)\rightarrow(h,c,t,p,k_w,k_s)$。

通过调用上述 API,能够将非结构化的文本转换为结构化的主题知识元,示例输出结果见表 8,其中文本主题 t 与关键句 k_s 取 Top 1 条;关键词 k_w 取 Top 5 条。最后,抽取 2 000 条文本调用该模型 API 进行人工校验,抽取合格的主题知识元的数量为 1 382 条,抽取的主题知识元合格率为 69.1%,其中主题知识元抽取合格,是指该主题知识元中所有元素均抽取正确,即文本主题、情感倾向、关键词与关键句元素均抽取正确时才视为抽取合格。

经过上述实验分析可得,本文提出的结构为“文本标题、文本内容、文本主题、主题情感倾向、主题关键词、主题关键句”的主题知识元,其内容相较于已有文献更加丰富;在主题抽取方面,有效地去除了冗余主题,并为知识元的其他元素抽取提供支撑;在主题情感分析方面,由于加入了主题特征作为输入,相较于单一 LSTM 模型,准确率提高了 2.3%;在关键词句方面,采用了主题相似度加权的抽取方法,相比单一的 TextRank 算法,其合格率提升 8%,同时使抽取的关键词句更加贴近文本主题。上述各实验结果表明了本文构建的主题知识元抽取方法是一种高质量的抽取方法。

5 总结与建议

本文提出了一种针对专业社交媒体的主题知识元抽取方法。首先通过 LDA 模型提取出专业社交媒体中文本的主题,并对主题进行聚类与去重,形成主题列表。其次,通过融合文本主题构建了适用于专业社交媒体本文的 T-LSTM 模型。然后,融合 TextRank 算法与主题相似度算法对文本中的关键词与关键句进行抽取,用于对主题与情感倾向的解释与补充。最后,对上述模型进行封装,通过封装程序将帖子文本转换为主题知识元,形成了完整的主题知识元抽取方案。

本文提出的模型能够较好的适应专业社交媒体论坛的文本特性,在主题提取方面进一步降低了主题的冗杂程度;在主题情感分析方面,围绕文本主题进行情感分析,提高了情感倾向分类的准确率;在关键词句方面,抽取得到的关键词句更加贴近文本主题。本文构建了完整的、系统的汽车社交媒体主题知识元的抽取方案,经过实验验证,抽取的主题知识元准确率到达 69.1%。此外,将深度学习与传统语义分析技术相结

合并引入到该主题知识元抽取方案中,是本抽取方法的特点。

专业社交媒体中主题知识元抽取的未来研究方向有:①对通过 LDA 模型输出的主题词用于构建本体,建立有层次结构与映射关系的主题体系,同时计算主题元素间的关联强度;②融合有监督或半监督的深度学习方法来抽取关键词句,从而提高主题知识元抽取的质量。

参考文献:

[1] 文庭孝, 侯经川, 龚蛟腾,等. 中文文本知识元的构建及其现实意义[J]. 中国图书馆学报, 2007, 33(6):91-95.

[2] 卜曲. 品牌社区网络结构及成员互动内容研究[J]. 现代商贸工业, 2016, 37(4):55-56.

[3] 吴婧. 试论网络论坛的文本构建特色[J]. 新闻研究导刊, 2016, 88(4):66-67.

[4] 王知津. 知识组织的目标与任务[J]. 情报理论与实践, 1999, 22(2):65-68.

[5] 温有奎, 温浩, 徐端颐,等. 基于知识元的文本知识标引[J]. 情报学报, 2006, 25(3):282-288.

[6] 姜永常. 知识构建的基本原理研究(下)——知识构建的技术支撑[J]. 图书情报工作, 2009, 53(6):100-104.

[7] 刘淼, 王宇. 基于主题句的期刊文献知识元库构建[J]. 情报杂志, 2012(11):145-149.

[8] 杨亮. 面向社交媒体的文本情感分析关键技术研究[D]. 大连:大连理工大学, 2016.

[9] YIN Y, SONG Y, ZHANG M. Document-level multi-aspect sentiment classification as machine comprehension[C]//PALMER M. Proceedings of the conference on empirical methods in natural language processing. Copenhagen: Association for Computational Linguistics, 2017:2044-2054.

[10] BLEI D, NG A, JORDAN M. Latent dirichlet allocation[J]. Journal of machine learning research, 2003(3):993-1022.

[11] 涂海丽, 唐晓波, 谢力. 基于在线评论的用户需求挖掘模型研究[J]. 情报学报, 2015, 34(10):1088-1097.

[12] ALEX G. Long short-term memory[M]//Supervised sequence labelling with recurrent neural networks. Berlin: Springer, 2012: 1735-1780.

[13] 梁军, 柴玉梅, 原慧斌,等. 基于极性转移和 LSTM 递归网络的情感分析[J]. 中文信息学报, 2015, 29(5):152-159.

[14] MIHALCEA R, TARAU P. TextRank: bringing order into texts [C]//RILL E. Proceedings of the conference on empirical methods in natural language processing. Barcelona: Association for Computational Linguistics, 2004:404-411.

[15] 韩龙士. 互联网+汽车新思维与商业模式创新[J]. 企业管理, 2015(7):104-106.

作者贡献说明:

林杰:确定论文结构,撰写论文;

苗润生:设计实验,进行数据收集与实验分析,撰写论文;

张振宇:进行数据整理,修改论文。

Research on Extraction Methods of Topic Knowledge Tuples in Professional Social Media

Lin Jie　Miao Runsheng　Zhang Zhenyu

School of Economics and Management, Tongji University, Shanghai 200092

Abstract: [**Purpose/significance**] Topic knowledge tuple is a knowledge unit for operating and managing knowledge oriented to knowledge themes. Accurately extracting topic knowledge tuples facilitates the storage, expression and retrieval of knowledge, and realizes knowledge creation and knowledge evaluation in the process of using knowledge. Therefore, this article discusses the existing extraction methods and then, by taking car products as an example, comes up with a method of extracting topic knowledge tuples from professional social media. [**Method/process**] First of all, this paper extracted a theme list from the users' comments in car forums with the LDA model. Secondly, based on the deep learning model T-LSTM which integrated thematic features, a sentiment analysis model suitable for the corpus of users in car forums was built. Then, by calculating the importance of each word in the TextRank diagram model and the similarity of each word's Word2Vec topic, we extracted key words and key sentences for the purpose of interpreting the extracted theme and sentiment orientation. Finally, the above methods were encapsulated into an integrated topic knowledge tuple extraction method. [**Result/conclusion**] In the experimental results, the qualification rate of extracted topic knowledge tuples reaches 69.1%. Experimental results show that the proposed method in this paper is capable of refining and extracting each element of knowledge tuples around the topic, meanwhile it can transforms unstructured information into structural knowledge.

Keywords: topic knowledge tuple　topic model　Long short-term memory (LSTM)　sentiment analysis